DOCUMENT RESUME

ED 415 696 FL 025 004

AUTHOR Brown, James Dean
TITLE Designing a Language Study.
PUB DATE 1997-00-00
NOTE 17p.; In: Classroom Teachers and Classroom Research; see FL 024 999.
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Research; Classroom Techniques; Ethics; *Research Design; *Research Methodology; *Sampling; Second Language Instruction; *Second Languages; *Validity

ABSTRACT

Some issues in the design of classroom research on second language teaching are discussed, with the intention of helping the researcher avoid conceptual pitfalls that may cripple the study later in the process. This begins with an examination of concerns in sampling, including definition of a population to be studied, alternative sampling strategies (random sampling, stratified random sampling), sample size, and the generalizability of the results based on the sample selected. Different types of variables (dependent, independent, moderator, control, intervening) and their roles in the research are then explained. A subsequent examination of research designs first defines treatment, control and experimental groups, and observations, and then distinguishes different designs, including true experimental, posttest-only, pretest-posttest (with and without control group), time series, and nonequivalent group designs. Characteristics, advantages, and problems with each design are noted. Validity is defined as the degree to which results can be accurately interpreted and effectively generalized, and these concepts are discussed further, including examination of threats to both internal and external validity. Finally, ethical issues are considered, basic guidelines are offered, and sources for further guidance are suggested. (Contains 26 references.) (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Gene Van Troyer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Chapter 5

ED 415 696

Designing a Language Study

James Dean Brown
University of Hawaii at Manoa

This paper introduces some of the overarching issues in second language research. They are issues which must be addressed before conducting a study so that the researcher can avoid conceptual pitfalls that may cripple the study later on. The discussion will begin with the considerations involved in sampling a group, or groups, of subjects to be used in a study. Next, the different types of variables that researchers define in a study will be covered. Then, some of the research designs that can be used in second language studies will be explored. In addition, the factors which may jeopardize the internal and external validity of language studies are covered. Finally, the ethical issues involved in collecting data, conducting research, and reporting the results will be discussed.

Sampling

In language studies, it is often necessary to use sampling techniques. To understand why such techniques are necessary, it is first important to grasp the difference between a population and a sample. In research, a *population* can usually be defined as the entire group of language speakers or learners that the researcher wants to study. Unfortunately, few researchers have the resources to study, for example, the entire population of ESL students studying ESL in American universities, or the entire population of EFL students in the world, or even all of the male chemistry students from Germany who are studying in the United States. As a result, most researchers prefer to use a *sample*, that is, a subgroup of the students

FL025364

BEST COPY AVAILABLE

representative of the given population. By using a sample, data can be practically and effectively collected, sorted, and organized. There are two basic strategies that are generally used in language studies for selecting samples from populations. These strategies are called random sampling and stratified random sampling. For both approaches, the purpose is to create an accurate sample, or subgroup, which can be said to be representative of the population as a whole.

Alternative Sampling Strategies

The underlying principle in *random sampling* is that each individual member of the population must have an equal chance of being selected into the sample. Three steps can be used to insure such equality of chance:

1. Clearly identify the population in which the researcher is interested.
2. Assign an identification number to each member of the population.
3. Choose the subjects for the sample on the basis of a table of random numbers.

A *table of random numbers* is a list of numbers, usually generated by a computer, that contains no systematic patterns. Most introductory statistics books contain such a list (for example, see Appendix A in Shavelson, 1981). Using a table of random numbers leaves the choices of who will be included in the sample up to a dispassionate and random table of numbers, rather than up to the researcher who may have subtle biases (conscious or unconscious) that could affect the results of the study. Once a large enough number of subjects is randomly selected, the resulting *random sample* can be assumed to be representative of the entire population from which they were drawn (Brown, 1988, pp. 111–113).

Other, more readily available techniques can be used to obtain a random sample. For example, the researcher might like to pull numbers out of a hat, use a deck of cards, or repeatedly throw a pair of dice in selecting subjects for a sample. Any technique wherein each member of the population has an equal chance of being selected, thereby ruling out biases on the part of the researcher, will be acceptable for random sampling, whether the sampling be for selecting subjects from a population for inclusion in a study, or for separating them into subgroups within the study itself.

Another strategy that is sometimes used in language studies is called *stratified random sampling*. In this case, four steps are usually used:

1. Clearly identify the population in which the researcher is interested.
2. Identify the salient characteristics of the population (called *strata*).
3. Randomly select members from each of the strata in the population (using a table of random numbers or other techniques described above).
4. Check to insure that the resulting sample has about the same proportions of each characteristic as the original population.

For instance, in the population of all ESL students studying at the University of Hawaii at Manoa (UHM), it might be useful to identify subgroups, or strata,

within the population based on the following characteristics: gender (male or female); country of origin; native language; academic status (graduate, undergraduate, or unclassified); and major (science, humanities, or undeclared). Given correct information about the proportions of these characteristics in the population of ESL students at UHM, the researcher could then randomly select from each of the strata in proportion to those population characteristics. The sample that results would intentionally take on the same proportional characteristics found in the entire population. Creating a stratified random sample still requires random sampling, but has the advantage of providing a certain degree of precision to the representativeness of the resulting sample—a fact which facilitates the use of the identified characteristics as variables in the study.

Decisions about which strategy (random or stratified random) to employ in a particular study must be reached rationally, and in advance. There are several considerations that must be kept in mind in making such decisions. First, it is generally useful to employ stratified random sampling when the population in question is fairly heterogeneous in nature. The concern is that random sampling might not provide for selection from each of the strata, or subgroups, in the population. Second, a stratified random sample becomes imperative when the samples involved will be small or the groupings within the study will be unequal in size. Third, it must be remembered that, if properly conducted, stratified random sampling has the advantage of letting the characteristics of the population determine which strata will be sampled. Hence, the stratified strategy is useful if the study will focus on the groups' characteristics as moderator or control variables (see Brown, 1988, pp. 11–18).

Alternatively, if the samples involved will be fairly large, straightforward random sampling can be employed. Random sampling is much easier to perform since there is no need to define the characteristics of the population. It is only necessary to assume that the sample represents the population from which it was taken. This assumption is widely accepted in research circles even though it is counter-intuitive for some language teaching professionals.

Sample Size

One of the first questions that will arise with regard to sampling is: How big must a sample be to be considered large enough? There is no easy answer to this question. However, it is clearly true that a large sample is better (in the sense of "more representative") than a small one. Consider a sample which includes all but 1% of a population of 1,000 language students (that is, a sample that contains 99% of the population). It is likely that such a sample is more representative of the population than one containing only 1% of it or 10% or 30%. However, knowing this does not answer the question of how big a sample must be to be considered large. Unfortunately, sample size decisions depend on the situation involved in the study as well as upon the types of statistics that will

BEST COPY AVAILABLE

be used. Statistics teachers will often give rules of thumb like the sample size should be at least 28 (or 30) per group or per variable. This is not bad advice *per se*, however, such rules of thumb are usually vague and imprecise, and in any case are conveying the minimum number that you will need for correctly applying many of the statistics that come up in research.

Another point of view is that, instead of estimating the bare minimum number of subjects, the researcher should be estimating the minimum number of subjects that would be necessary for a statistically significant result to be obtained (if it really exists in the population) given the application of a particular statistical procedure under the conditions of the study that is being planned. Such estimations can be made by using *power analysis*. One thing that power analysis can be used for is to analyze the relationship between the probability of finding a statistically significant result and the sample size given a particular set of expected results. If, for instance, a researcher wanted to estimate the number of subjects that would be necessary to find a statistically significant difference between the means of two groups of subjects, it could be done mathematically on the basis of pilot data, or other previous research that may be available in the literature. Such estimates can be made for a variety of the statistical procedures used for mean comparisons, correlation, and regression, as well as comparisons of frequencies (for more on power analysis, see Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990). Unfortunately, power analysis is mathematically complex. However, there is computer software available (e.g., Borenstein & Cohen, 1988) that can resolve this problem.

In short, when thinking about sample size, the best strategy is to make sure that the population is clearly defined, and that the sampling procedures make sense. If pilot data or other previous research is available, it will prove helpful to use power analysis to estimate the sample size that is necessary to find a significant effect if it exists. If pilot data are not available, you may have to design your study such that the samples involved "seem" large enough to be representative, while keeping in mind that a good rule of thumb is the larger the sample size the better. The issues involved in sampling are somewhat subjective, and must in part be left up to the researcher. Sampling procedures are important partly because of the way that they affect the generalizability of the study.

The *generalizability* of a study can be defined as the degree to which the results are meaningful beyond the study itself with regard to the entire population in question. If the sampling techniques have been properly conducted and the sample is large enough, there should be no question in the researcher's mind (or in a reader's mind) as to the degree to which the sample represents the population. If there is some question, then the sampling techniques should be improved or the sample sizes increased, or both. (For more information on sampling and its use in language studies see Brown, 1988; Hatch & Lazaraton, 1991.)

Different Types of Variables

A *variable* is anything that can vary in a study. However, research is largely the study of what happens when variables are systematically manipulated in planned combinations. There are essentially five roles that variables can play in a study: dependent variables, independent variables, moderator variables, control variables, and intervening variables.

The *dependent variable* in a study is the variable of primary focus. It can also be thought of as the variable that is measured and studied to determine if other variables have an effect on it, or are related to it. The *independent variable* in a study is the variable that has been selected by the researcher in order to study its effect on the dependent variable (hence, the independent variable is sometimes also called the manipulated variable). For instance, for a research question like "What is the effect of X on Y?", X is the independent variable and Y is the dependent variable. Or, for a research question like "How well does X predict Y?", X is the independent variable and Y is the dependent variable.

The relationship between the independent and dependent variables is central to any study. However, sometimes the researcher will also want to include a *moderator variable* in order to determine the effect of the moderator variable on the relationship between the dependent and independent variables. Thus, if a moderator variable were included, a question like the following could be posed: "What is the effect of X on Y when Z is present or absent?" In this last case, X is the independent variable, Y is the dependent variable, and Z is a moderator variable.

In language research, there are usually variables other than the dependent, independent, and moderator variables which cannot be included in the design or otherwise directly studied. Nonetheless, these variables must be accounted for often as control variables. *Control variables* are variables which are eliminated from the study, held constant, or otherwise kept from interfering with the study of the central relationship between the independent and dependent variables. For instance, in a study of the effect of Method A on English language proficiency (as measured by TOEFL), the researcher might compare the TOEFL scores of two groups, one who had been taught by Method A and another who had received no instruction, a control group. The researcher would be most interested in the relationship between the independent variable, Method, and the dependent variable, English Language Proficiency. However, there are a number of variables which might interfere with the relationship between Method and English Language Proficiency: gender, intelligence, aptitude, years of language study, etc. The researcher might choose to control gender by eliminating all males from the study. The researcher might further choose to use only students who had studied six years of English to hold the years-of-study variable constant. Random selection can also be used to create groups that are theoretically equal on all variables except those being manipulated as independent, dependent, and moderator variables.

Perhaps, the most confusion is caused by the term "intervening" variable because it is used in two distinctly different ways. On the one hand, *intervening variable* is used to describe the construct which underlies the relationship between the independent and dependent variables. For instance, in the example study above on the effect of Method A on English Language Proficiency, the researcher might label the construct underlying the effect as "method effect" or "learning" or "language acquisition" depending on how it is conceptualized.

On the other hand, *intervening variable* is used to describe a variable that is unanticipated in a study, yet surfaces as being a possible explanation for the relationship between the independent and dependent variable. In the example study, it might turn out that any difference discovered in the proficiency scores of the Method A and Method B groups were caused by an unanticipated intervening variable rather than by the methods themselves. For instance, it might turn out that the teacher of the Method A class was just a better teacher than the teacher of the control group. Thus, a teacher effect turns out to be a possible intervening variable in the sense that it was unanticipated yet has potential explanatory power.

Research Designs

To understand the basic designs that are used in quantitative language studies, it will first be necessary to define some of the fundamental terms that are used. The first idea that must be understood is that of a *treatment*. A treatment is something that the experimenter does to one group so as to study the effects of the treatment on the people involved. A treatment may be a specific teaching strategy, application of a set of materials, use of a particular reward system, or any other experience that the researcher wants to apply to the subjects for the study. Typically, for the sake of comparison, one group receives the treatment while another group does not. Thus, the subjects are divided into two or more groups: a control group and one or more experimental groups. The *control group* usually receives no treatment, or a placebo (some substitute that is predicted to have no effect), while the *experimental group* receives the treatment. In a language program, the treatment is likely to be some aspect of the language teaching or learning experience.

The reason for administering a treatment to the experimental group and nothing to the control group is to determine whether the treatment has had an effect. In order to do so, one or more observations must occur which allow for comparisons of the two types of groups. These *observations* may take many forms. In quantitative studies, observations may be simple tallies, rankings, or test scores. The point in making observations is that something of interest to the researcher must be observed or measured so that comparisons can be made between the control and experimental groups. Naturally, whatever is observed

or measured must be related to the treatment. Thus, in a language program, if the treatment was some form of pedagogy, you might be interested in observing the language achievement test scores in order to determine the effect of the treatment on achievement.

It is important to note that studies involving anything other than examination and description of test scores are difficult to conduct. The study may be designed in an airtight manner (difficult in any teaching or learning situation), but in addition, considerable knowledge of statistics must be applied—usually more than the knowledge provided in one or two statistics courses. This warning is meant to encourage budding researchers to seek adequate guidance in designing quantitative studies and analyzing the statistical results.

The two sections that follow will explain two categories of quantitative studies: true experimental designs and quasi-experimental ones. This is a very useful distinction explained much more fully in Campbell and Stanley (1963).

True experimental designs are the most controlled language studies. They must be carefully planned from beginning to end. Hence, they are the closest thing in language studies to what most teachers believe scientific experiments are like. One of the keys to identifying a true experimental design is that the subjects in the study must be randomly selected from the population being studied, and randomly assigned to the treatment or control group. ***Randomly*** is used here strictly in the sense that it was defined above. As described in the earlier section on *Sampling*, this must be done so that every member of the population has an equal chance of being selected. If these procedures are followed and the resulting groups are large enough, the researcher is justified in assuming that the two groups have very much the same characteristics. Thus, true experimental designs have random selection as a precondition. The same thing is true for posttest only designs, pretest-posttest designs, or any combination of the two.

The *posttest-only design* (one type of true experimental design) is particularly dependent on random selection because it is assumed on the basis of sampling theory that the experimental and control groups are equivalent at the outset of the study. Such a study is designed as shown in Figure 1. Notice that step one is to use random selection to create equivalent groups. The experimental group receives the treatment, while the control group does not (or receives a placebo). Both groups are then observed on the same scale and the performances of the two groups are compared. If the experimental group has significantly higher performance than the control group, arguments can then be built that the treatment has an effect. The degree to which such claims can be made will naturally depend on the magnitude of the differences in performance.

The *pretest-posttest design*, while it also assumes random selection of the two groups, allows the researcher to check the equivalence of the two groups at the beginning of the study, usually a pretest of some sort. Such a study would typically be laid out as shown in Figure 2. This additional step allows for checking the

BEST COPY AVAILABLE

Figure 1
True Experimental Design, Posttest Only

Step 1		Step 2		Step 3	Step 4
Random Experimental Group	→	Receives Treatment	→	Observation	Performances Compared
Random Control Group	→	Receives No Treatment	→	Observation	

equivalence of the two groups in Step 2, but also allows for studying the amount of gain that has been made by each group between Steps 2 and 4. This potential for studying gain allows the researcher to consider additional issues. For example, if there is a difference between the two groups on the posttest, the researcher can study whether the difference is as large as the difference between the pretest and posttest performances of the experimental group. If this is not true, the observed differences may have some source other than, or additional to, the treatment. Thus, the pretest-posttest design is generally more powerful than the posttest only design because more inferences can be drawn. Pretest-posttest designs can become much more complex including various types of treatments used simultaneously and various observation techniques used in the same study.

Figure 2
True Experimental Design, Pretest and Posttest

Step 1		Step 2		Step 3		Step 4	Step 5
Random Experimental Group	→	Observation (Pretest)	→	Receives Treatment	→	Observation (Posttest)	All Observations Analyzed
Random Control Group	→	Observation (Pretest)	→	Receives No Treatment	→	Observation (Posttest)	

From a practical point of view, true experimental designs are often doomed in real language teaching settings. First, students are rarely randomly selected. Thus, many researchers are working either with what is called an *intact group* or with the entire population of students when they set out to do a study. Second, the researcher cannot set aside half of the students, randomly selected or otherwise, to receive no language training, or a placebo. Either students want the training or they do not, and language researchers are seldom in the moral or monetary position to simply withhold treatment (training) from one half while the other half receives training. As a result, language researchers are more likely to turn to what is called a quasi-experimental design. *Quasi-experimental designs*, though less than perfectly controlled, provide useful alternatives to true experimental designs. Quasi-experimental designs are adequate for the purposes of studying many language issues—particularly if no sweeping claims are going to be generalized from the results. According to Campbell and Stanley (1963), the main characteristic that makes a quasi-experimental design more practical for language studies is that the researcher has more control over the collection of data in terms of scheduling and deciding who will participate. However, the designs are weaker, and the results must be interpreted very carefully. Three types of quasi-experimental designs will be presented here: pretest-posttest designs (without control group), time series designs (without control group), and nonequivalent groups designs.

The *pretest-posttest design without control group* is like the pretest-posttest design discussed above except that it lacks a control group. Such a design is shown in Figure 3. This type of design could be used as follows: A general proficiency pretest could be given at the beginning of a language program (the treatment) and again as a posttest at the end of the program. If there is a large gain in average scores between the beginning and end of the program, it might be judged as a success. However, because there is no control group in such a study, the researcher can never know for sure that the gain was not a result of language exposure outside of the program, or a result of a testing effect (that is, the effect of having taken the test twice), or a result of some other undetermined factor. In other words, the observed gains in scores may have been due to factors other than the learning that took place in the program.

Figure 3
Quasi-Experimental Design, Pretest and Posttest

Step 1	Step 2	Step 3	Step 4	Step 5
Experimental Group	→ Observation (Pretest)	→ Receives Treatment	→ Observation (Posttest)	] Observations Compared

BEST COPY AVAILABLE

Time series designs are more elaborate versions of the pretest-posttest design. The only striking difference is that, in lieu of one pretest and one posttest, a series of observations, or tests, are made. Then, a treatment is inserted in the middle of this series. Such a design is described in Figure 4 (in which "O" stands for Observation). In a time series design, the researcher can claim that the potential consequences of the testing effect mentioned above are controlled in that all students are made thoroughly familiar with the format and content types on the observation instruments long before the treatment comes into the picture. One problem that arises with this type of design is that it sometimes calls for the development of numerous instruments, all of which must be very similar in what they measure.

Figure 4
Quasi-Experimental Design, Time Series

O	→	O	→	O	→	O	→	Treatment	O	→	O	→	O	→	O	All
1		2		3		4			5		6		7		8	Observations
																Analyzed

The *nonequivalent groups design* is different from the true experimental pretest-posttest design only in that the subjects are not randomly selected into the experimental and control groups. Such a design is shown in Figure 5. Because the groups are not randomly selected, they cannot be assumed to be equivalent at the beginning of the study. As a result, the equivalence of the groups must be checked in Step 2 (or otherwise controlled statistically). If it is possible to set up a control group in this manner and the groups do indeed prove to be equivalent at the beginning of the study, the nonequivalent groups design can prove fairly powerful. However, if such a control group cannot be established, the quasi-experimental version of the pretest-posttest design (Figure 3) may be the most effective design that can be used.

Figure 5
Quasi-Experimental Design, Nonequivalent Groups

Step 1		Step 2		Step 3		Step 4	Step 5
Experimental Group	→	Observation (Pretest)	→	Receives Treatment	→	Observation (Posttest)	All Observations Analyzed
Control Group	→	Observation (Pretest)	→	Receives No Treatment	→	Observation (Posttest)	

There are many other types of complex designs (see Campbell & Stanley, 1963, or Tuckman, 1978), and numerous ways of grouping and analyzing the results of those designs (see for instance, Keppel, 1973; Kirk, 1968; Pedhazur, 1982; Tabachnick & Fidell, 1989).

Validity

The validity of a study can be defined as the degree to which the results can be accurately interpreted and effectively generalized. The first part of this definition—the degree to which the results can be accurately interpreted—is often referred to as *internal validity*. The second part, the degree to which the results can be generalized, is often labeled *external validity*. Table 1 lists the different factors that can affect the validity of a study (after Campbell & Stanley, 1963).

Internal Validity

The eight threats to internal validity, listed above, are variables that must be controlled in designing a study so that the results can be accurately interpreted.

History includes anything that happens to the subjects, other than the intended treatment, between the observations in a study. For example, for the design shown in Table 1, history would be anything, other than the treatment,

Table 1
Factors Threatening Validity (after Cambell & Stanley, 1963)

Type of Validity Factor
Internal Validity
1. History
2. Maturation
3. Testing
4. Instrumentation
5. Statistical regression
6. Selection bias
7. Experimental mortality
8. Selection-maturation interaction
External Validity
9. Reactive effects of testing
10. Interaction of selection biases and the treatment
11. Reactive effects of experimental arrangements
12. Multiple treatment interference

BEST COPY AVAILABLE

that occurs between the pretest and posttest for either the experimental or control group in the True Experimental Design, Pretest, and Posttest.

Maturation refers to any of the processes in the subjects' lives that occur because of the passage of time and might interfere with interpretation of the results of a study. For instance, fatigue, hunger, aging, changing schools, or passage through puberty would all be maturation factors that the researcher might want to consider.

Testing describes any influence that taking one test has on the scores of another test. For instance, taking the pretest shown in Table 1 might affect the scores on the posttest. The testing effect might be particularly pronounced if the type of test involved were completely new to the subjects involved. Consider a group of subjects who had never taken a cloze test before. If one were administered as a pretest, the subjects might learn test taking strategies that would make them more comfortable and make them score higher on a subsequent posttest, regardless of any treatment that was administered.

Instrumentation involves the impact of variations in the tools of measurement or problems with the reliability of those tools (for much more on this latter topic, see Brown, 1995a, 1995b) on the obtained measurements, or scores. For example, a problem of instrumentation would arise if version A of a test was used in the pretest, but version B was used on the posttest. The problem is that any differences in performance could be due to discrepancies in the versions of the test (the instruments) rather than to any treatment involved.

Statistical regression describes the moderating effects of selecting groups with extreme scores, either very high, very low, or both. Under such conditions, the probability is that students with high scores will tend to score lower (i.e., closer to the average score), while students with very low scores will tend to score higher (i.e., closer to the average score) for reasons having nothing to do with any treatments involved.

Selection bias describes the impact of selecting subjects into the groups of a study for reasons other than chance. For instance, if the subjects for the experimental group in Table 1 were selected from students in 8:00 a.m. ESL classes, while the subjects in the control group were selected from students in 4:00 p.m. ESL classes, there might be differences in the groups based on class time preference that have nothing to do with the treatment involved.

Experimental mortality refers to the influence of subjects dropping out of one or more of the groups in a study. For example, in Table 1, some subjects in the control might be present for the pretest but absent for the posttest. These absences might cause differences in the results that had nothing to do with the treatment.

Selection-maturation interaction describes the effect of the maturation and selection bias variables (defined above) acting together.

External Validity

The four threats to external validity listed above can affect the generalizability of the results.

Reactive effects of testing and treatment describe the influence of taking a pretest on the sensitivity of the subjects to the treatment. For instance, if the treatment involved the use of cloze tests to practice reading prediction and the pretest was also a cloze test, the pretest might affect the subjects' sensitivity to the treatment. In other words, the generalizability of the results might be in question because the results depend on the use of a particular test.

Interaction of selection biases and the treatment. If there is some relationship between the group from which the subjects were selected and the effects of the treatment, interactions are said to exist between selection biases and the treatment variable. In other words, in any study, there is the possibility that any effects that are found are only true for the population from which the groups were selected. It is also possible that the characteristics of that particular population may cause the treatment to be effective where it would not be in another population. In such a situation the selection bias would be interacting with the treatment and thus affecting the generalizability of the results.

Reactive effects of experimental arrangements. This refers to the impact of the fact that the treatment was applied under experimental conditions rather than real world conditions. For example, some pedagogical techniques might appear to work very well as a treatment under classroom conditions but have no correspondingly beneficial effect on the students' use of the language in the real world. The generalizability of the results to the real world would be in question.

Multiple treatment interference refers to the effects of applying more than one treatment to the same subjects. Under these conditions, the effects of one treatment cannot be disentangled from the effects of others, and thus the results cannot be generalized to situations that do not contain the multiple treatments.

Multiple Threats to Validity

Unfortunately, threats to validity in research are seldom as simple as Table 1 would suggest. This is because there may be numerous threats to validity operating at the same time. Since the overall confusion caused by simultaneously having multiple threats to the validity of a study may well be synergistic, it is crucial that researchers guard against and control any and all of these problem factors, preferably while planning a study. (For more information on factors that threaten the validity of a study and how to control them, see Brown, 1988; Campbell & Stanley, 1963; Hatch & Lazaraton, 1991; Tuckman, 1978.)

Ethics

Ethics in social science research have been considered from a number of perspectives. For an overview of this work see Kimmel (1988). Over the years, various organizations associated with social sciences research have provided guidelines for their memberships. For example, the American Psychological

Association has provided various kinds of guidelines for the ethical conduct of research (American Psychological Association, 1953, 1981, 1982, & 1985). According to Kimmel (1988), ethical problems in social sciences research may have a number of the following characteristics:

1. The complexity of a single research problem can give rise to multiple questions of proper behavior.
2. Sensitivity to ethical issues is necessary but not sufficient for solving them.
3. Ethical problems are the results of conflicting values.
4. Ethical problems can relate to both the subject matter of the research and the conduct of the research.
5. An adequate understanding of an ethical problem sometimes requires a broad perspective based on the consequences of research.
6. Ethical problems involve both personal and professional elements.
7. Ethical problems can pertain to science (as a body of knowledge) and to research (conducted in such a way as to protect the rights of society and research participants.
8. Judgments about proper conduct lie on a continuum ranging from the clearly unethical to the clearly ethical.
9. An ethical problem can be encountered as a result of a decision to conduct a particular study or a decision not to conduct the study.

Commandments

In language related research, some of the most important ethical and professional issues might best be summed up by ten straightforward commandments (adapted from Brown 1984). These commandments cover the researcher's ethical and professional responsibilities with regard to the participants, analyses, and audience of a study:

Participants

I. Thou shalt not abuse thy subjects in any manner including abuses of their persons, time, or effort, and thou shalt obtain thy subjects' informed consent if required by thy institution.

II. Thou shalt not abuse thy colleagues by collecting data from their students without permission, or by using too much precious class time.

III. Thou shalt reward thy subjects' and colleagues' efforts at least by giving them feedback or information on what happened in the study.

Analyses

IV. Thou shalt guard against consciously or subconsciously modifying thy data so that the results support thy views and prejudices.

V. Thou shalt select the appropriate statistical tests.

VI. Thou shalt check the assumptions that underlie all statistical tests.

Audience

VII. Thou shalt explain thy research clearly so that it can be understood by thy readers.

VIII. Thou shalt organize thy report using conventional sections, headings, and other conventions (see American Psychological Association, 1994) so that thy readers can easily follow thy study.

IX. Thou shalt interpret thy results carefully guarding against the temptation to over-interpret, or generalize beyond that which thy results warrant.

Above All Else

X. Thou shalt continue to learn, read, and grow as a researcher so that thou can better serve thy field.

Stating the ethical issues in the form of commandments may at first seem to be intended as tongue-in-cheek humor, but these are not to be taken lightly. Indeed, the entire enterprise of research in language studies hinges on cooperation between subjects, colleagues, researchers, and readers. Researchers should avoid contributing to the already abundant negative feelings about statistical research.

Conclusion

This paper began with a discussion of the issues involved in sampling a group, or groups, of subjects to be used in a study. Then, data collection instruments were examined in terms of the four scales of measurement that can be used. Next, a number of research designs were explored. Then, the factors which may jeopardize the internal and external validity of language studies were surveyed. Finally, the ethical issues involved in collecting data, conducting research, and reporting the results were covered. In short, a great many crucial issues have been covered in this paper—issues that must be thought through before conducting a study. A little effort spent in the planning stages of a study can save the enormous amount of energy necessary to recover if things begin to come unraveled after the study has begun.

References

American Psychological Association. (1953). *Ethical standards of psychologists*. Washington, DC: American Psychological Association.

American Psychological Association. (1981). *Ethical principles of psychologists*. Washington, DC: American Psychological Association.

American Psychological Association. (1982). *Ethical principles in the conduct of research with human participants*. Washington, DC: American Psychological Association.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (1985). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program*. Hillsdale, NJ: Lawrence Erlbaum.

Brown, J. D. (1983). An exploration of morpheme-group interactions. In K. M. Bailey, H. Long, & S. Peck (Eds.). *Second language acquisition studies*. Cambridge, MA: Newbury House.

Brown, J. D. (1984). Moral, ethical and social considerations in quantitative TESOL research. Paper presented at the TESOL Convention, Houston, TX.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. London: Cambridge University.

Brown, J. D. (1995a). *The elements of language curriculum: A systematic approach to program development*. Boston, MA: Heinle & Heinle.

Brown, J. D. (1995b). *Testing in language programs*. Eaglewood Cliffs, NJ: Prentice Hall.

Brown, J. D., Knowles, M., Murray, D., Neu, J., & Violand-Hainer, E. (1992). *A survey of research and practice in the TESOL organisation*. Paper presented at the TESOL Convention, Vancouver, Canada.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Keppel, G. (1973). *Design and analysis: A researcher's handbook* Englewood Cliffs, NJ: Prentice Hall.

Kimmel, A. J. (1988). *Ethics and values in applied social research*. Newbury Park, CA: Sage.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences. Belmont*, CA: Brooks/Cole.

Kraemer, H. C. & Thiemann, S. (1987). *How many subjects?* Newbury Park, CA: Sage.

Krashen, S. (1977). Some issues relating to the monitor model. On *TESOL '77*. Selected papers from the 11th annual TESOL convention, Miami (pp. 144–158). Washington, DC: TESOL.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*ed Newbury Park, CA: Sage.

Pedhazur, E. J. (1982). *Multiple regression in behavioural research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.

Shavelson, R. (1981). *Statistical reasoning for the behavioural sciences*. Boston: Allyn and Bacon.

Tabachnick, B. G. & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper Collins.

Tuckman, B. W. (1978). *Conducting educational research*. New York: Harcourt Brace Jovanovich.